D04 Core Cornetto

Date: 07 May 2008

Project: STE05039
http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm

Funded by the Stevin framework (www.taalunieversum.org/stevin)

Authors:

- Piek Vossen, Irion Technologies/ Faculteit der Letteren, Vrije Universiteit Amsterdam
- Isa Maks, Faculteit der Letteren, Vrije Universiteit Amsterdam
- Roxane Segers, Faculteit der Letteren, Vrije Universiteit Amsterdam
- Hennie van der Vliet, Faculteit der Letteren, Vrije Universiteit Amsterdam

1. Introduction

The Cornetto project (STE05039) is funded by the Nederlandse Taalunie in the STEVIN framework (www.taalunieversum.org/stevin). The goal was to build a lexical semantic database for Dutch, covering 40K entries, including the most generic and central part of the language. Cornetto combines the structures of both the Princeton Wordnet and FrameNet for English, by combining and aligning two existing semantic resources for Dutch: the Dutch wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Martin et al 1999). For more information on the Cornetto database architecture see deliverable *D01 Design of the Cornetto database*.
To create the initial Cornetto database, the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch WordNet (DWN) have been automatically aligned. The final Cornetto Database can be divided into two quality subsets: the automatically aligned part with reliable quality scores for the linking between a lexical unit and a synset, and the part that has been manually checked, edited and confirmed after the automatic alignment. This latter high-quality part of Cornetto, covers at least 10K of word meanings and will be referred to by the name Core Cornetto. In this deliverable, we describe the work and decisions made for the automatic alignment of the database, and the selection and work on Core Cornetto.

2. Automatical alignment of the database: automatic quality labelling

The first step in creating the Cornetto database was the automatic alignment of the word meanings in the Referentie Bestand Nederlands (RBN) and the Dutch WordNet (DWN) based on eight strategies:

- a word has one meaning and no synonyms in both RBN and DWN
- a word has one meaning in both RBN and DWN
- a word has one meaning in RBN and more than one meaning in DWN
- a word has one meaning in DWN and more in RBN
- if the broader term (BT) of a set of words is linked, all words which are under that BT in the semantic hierarchy and which have the same form are linked
- if some narrow term (NT) in the semantic hierarchy is related, siblings of that NT that have the same form are also linked.
- word meanings that have a linked domain, are linked
- word meanings with definitions in which one in every three words is the same (there must be more than one match) are linked.

Each of these heuristics will result in a score for all possible mappings between word meanings. The number of links found per strategy is shown in Table 1.To weigh the heuristics, we manually evaluated each heuristics. Of the results of each strategy, a sample was made of 100 records. Each sample was checked by 8 persons (6 staff and 2 students). For each record, the word form, part-of-speech and the definition was shown for both RBN and DWN (taken from VLIS). The testers had to determine whether the definitions described the same meaning of the word or not. The results of the tests were averaged, resulting in

a percentage of items which were considered good links. The averages per strategy are shown in Table 1.

	Conf.	Dev.	Factor	LINKS	
1: 1 RBN & 1 DWN meaning, no synonyms	97.1	4,9	3	9936	8,1%
2: 1 RBN & 1 DWN meaning	88.5	8,6	3	25366	20,8%
3: 1 RBN & >1 DWN meaning	53.9	8,1	1	22892	18,7%
4: >1 RBN & 1 DWN meaning	68.2	17,2	1	1357	1,1%
5: overlapping hyperonym word	85.3	23,3	2	7305	6,0%
6: overlapping hyponyms	74.6	22,1	2	21691	17,7%
7: overlapping domain-clusters	70.2	15,5	2	11008	9,0%
8: overlapping definition words	91.6	7,8	3	22664	18,5%

The minimal precision is 53.9 and the highest precision is 97.1. Fortunately, the low precision heuristics also have a low recall. On the basis of these results, the strategies were ranked: some were considered very good, some were considered average, and some were considered relatively poor. The ranking factors per strategy are:

- Strategies 1, 2 and 8 get factor 3
- Strategies 5, 6 and 7 get factor 2
- Strategies 3 and 4 get factor 1

A factor 3 means that it counts 3 times as strong as factor 1. It is thus considered to be a better indication of a link than factor 2 and factor 1, where factor 1 is the weakest score. The ranking factor is used to determine the score of a link. The score of the link is determined by the number of strategies that apply and the ranking factor of the strategies. Within the Cornetto database, this score thus is an indicator of the quality of a link between a synset and a lexical unit. In the case of the word *thee* (tea), the following alignment scores are presented:

```
thee:1, NOUN, LU=r_n-37550, SY=d_n-25670; 91.6, -C
thee:2, NOUN, LU=r_n-37551, SY=d_n-31286; 83.1, -C
thee:3, NOUN, LU=r_n-37552, SY=; -1.0, -C
thee:4, NOUN, LU=d_n-417507, SY=d_n-29386; -1.0, -C
```

This means that *tea* has three meanings in RBN (see the identification numbers starting with a ‘r’) and three meanings in DWN (see the identification numbers starting with a ‘d’). It now has four meanings in Cornetto, two of which are aligned and two are not. The first two generated mappings have a high score that suggests the probability of a correct mapping. The third meaning only exists

in RBN and has no corresponding synset; the score therefore is -1. The fourth meaning only consists in DWN and could not be mapped to an existing lexical unit. Therefore, an extra empty LU was created (d_n-417507) with a score -1, suggesting that this is a dummy mapping. The '-C' at the end of each line indicates that these mappings have not yet been manually edited and confirmed.

After the manual alignment (see next paragraph), the alignment scores will be replaced with a 'MC' that indicates that each meaning has been edited if necessary and that the mapping was manually confirmed. Note that for the third meaning in RBN a new synset was created (c_592) and that the automatic generated lexical unit was replaced with a new and edited one (c_545867).

```
[*] thee:1, NOUN, LU=r_n-37550, SY=d_n-25670;, MC
[*] thee:2, NOUN, LU=r_n-37551, SY=d_n-31286;, MC
thee:3, NOUN, LU=r_n-37552, SY=c_592;, MC
thee:4, NOUN, LU=c_545867, SY=d_n-29386;, MC
```

In total, 136K linking records are stored in the Cornetto database. Within the database, only the highest scoring links are used to connect RBN LUs to DWN LUs and likewise to synsets. There are 58K top-scoring links, representing 41K word meanings. In total 47K different RBN word meanings were linked, and 48K different DWN word meanings. 19K word meanings from RBN were not linked, as well as 59K word meanings from DWN.

3. Manual Alignment

The automatic alignment was followed by the manual alignment that focuses on the important words in the database: high-frequent words and high-polysemous words. These words are also most difficult to align which follows from low-scoring of links and absence of links.

We defined four groups of expected problematic cases on which we concentrated the manual work:

? 1800 most polysemous nouns (4 or more meanings)
? 1000 frequent nouns with a semantic shift
? 1000 frequent verbs
? 250 most frequent adjectives

Polysemous nouns

The list of 1800 most polysemous nouns was selected after the conclusion that the automatic alignment was not always reliable. By editing the most polysemous words, we also had the chance to correct one of the most crucial parts of the Cornetto database since high polysemy tends to go hand in hand with high frequency, complex morpho-syntax and complex semantics. We selected all nouns with four ore more meanings (table 2). Note that this table represents the initial polysemy after the automatic alignment; if there were no matches across the two databases, an empty sense was created. In the final report some statistics will show the difference in the polysemy before and after the manual work. This could not be represented in this deliverable, since the manual alignment is still ongoing.

Nr. of senses after alignment	Nr. words (Nouns)
10 and more	62
9	41
8	75
7	126
6	212
5	389
4	1026
3	2507
2	7293
1	64298

Table 2: Polysemy distribution of nouns

Frequent nouns with a semantic shift

The RBN uses the semantic shift for groups of words that show the same semantic behaviour. Because of the difference in approach, the DWN resource will have an extra synset for the meaning that is implied with a shift in the LU. There are about 30 different defined types of shifts like Dynamic ? Non-dynamic and Nonhuman ? Concrete.

We found that the matching of LUs from RBN to synonyms in DWN was more likely to be incorrect for all words labeled with a shift in RBN. For this reason a part of the manual editing was concentrated on 1000 frequent nouns with a semantic shift.

The need for editing these shift cases can be illustrated by the word *bekendmaking* (*announcement*) that has one LU with a shift in RBN from Dynamic to Non-dynamic. This means that (in Dutch) an *announcement* can be a process and the result of this process. In DWN, we find a synset for each of these aspects, stating that the first one is a subclass of the SUMO term 'Communicating', and the second one is equivalent to 'Statement'. We have used this ontological evidence as an argument to split the LU and define the difference between the two meanings in terms of the definition and the semantic relations as is shown in Fig. 1:

Dynamic X	*announcement*
LU resume	'the announcing'
LU combinatorics/example	-
HAS_HYPERONYM	statement (*dyn. in Dutch*)
XPOS_NEAR_ SYN	announcing
SUMO	+ Communicating

Non-dynamic X	*announcement*
LU resume	'something announced'
LU combinatorics/example	-
HAS_HYPERONYM	message
ROLE_RESULT	announcing
SUMO	+ Statement

Figure 2: example of splitting a LU with a semantic shift

Frequent verbs

The low-scoring links within the group of verb synsets and lexical units are in great deal due to the difference regarding the underlying principles of meaning discrimination which plays an important role in the alignment of synsets and lexical units.

As long as there is a one-to-one mapping from LUs and synsets, the features of the two resources will probably match. Difficulties arise however when the mapping is not one-to-one. Frequent verbs are often very polysemous. The RBN, as the source of the LUs, tries to deal with polysemy in a systematic and efficient way. The synsets, however, are much more detailed on different readings. As a result, in many cases there are more synsets than LUs. In combination with the detailed information on complementation, event structure and lexical relation, this results in interesting editing problems.

A typical example of an economically created LU in combination with a detailed synset is *aflopen* (*to come to an end, to go off (an alarm bell)*, *to flow down, to run down, to slope down, etc.*). Input to the alignment are seven LUs and 13 synsets. Much of the asymmetry was caused by the fact that one of the LUs represents one basic and comprehensive meaning: *to walk to, to walk from , to walk alongside something or someone*. In DWN these are all different meanings, with different synsets. This is the result of describing lexical meaning by synsets; these three readings of *aflopen* obviously have a lot in common, but they match with different synonyms. Aligning the LU's and synsets leads to splitting the LU's and may lead to subtle changes in the complementation patterns, event structure and certainly to adapting and extending the combinatorical information. Sometimes the LUs are more detailed. In that case a synset must be split, which of course gives rise to changes in all related synsets and to new sets of lexical relations.

Most frequent adjectives

A considerable part of the adjectives was not successfully aligned by the automatic alignment procedures. This was especially due to the fact that adjective synsets have few semantic relations lacking hyperonyms and hyponyms. By consequence, the automatic alignment strategies which involve broader and narrower terms are in these cases not applicable.

Another problematic aspect of the adjective synsets is the fact that the automatically formed DWN adjective synsets have not – unlike the noun and verb synsets – been edited and corrected manually before. To be able to deal in a systematic way with these problems, we introduced the use of a semantic classification system for adjectives (Hundschnurser & Splett, Germanet). For more details see: *D03 Top level ontology, relation constraints and assignments*

The total of 3250 words in these four lists equal to approximately 10.000 word meanings and cover the most frequent part of the Dutch language. Due to the different problems the focus of each editing strategy was slightly different, but for each group four basic steps were formulated:

- Checking and editing the mapping of RBN and DWN
- Checking and editing the mapping of DWN to English Wordnet 2.0
- Checking and editing the mapping of DWN to Wordnet Domain labels
- Checking and editing the mapping of DWN to SUMO and MILO

Each of these steps can be divided in several actions:

Checking and editing the mapping of RBN and DWN

- ? Creating new synsets and lexical units
- ? Splitting of lexical units in two different meanings (mostly done in lexical units with a semantic shift)
- ? Splitting of synsets in two or more different meanings (mostly done in adjective synsets)
- ? Adding semantic relations between new and existing synsets
- ? Adding examples in the lexical units

Checking and editing the mapping of DWN to English Wordnet 2.0

- ? Selecting the correct equivalence relations in terms of e.g. a near-synonym relation, a near hypernym relation, etc.

Checking and editing the mapping of DWN to Wordnet Domain labels

- ? Selecting the correct Wordnet Domain labels

Checking and editing the mapping of DWN to SUMO and MILO

- ? Selecting SUMO terms by e.g equivalence and subsumption.
- ? For some cases the mapping with SUMO was problematic; therefore we created the possibility to extend the ontology by creating triplets based on first order logic. For more information on assigning ontology terms see deliverable *D03 Top level ontology, relation constraints and assignments*.

4. Additional Work

In addition to the manual alignment and editing of Core Cornetto we have planned some work and strategies to improve the Cornetto database.

4.1 Database improvement

The main points are the improvement of the assigned SUMO terms by traversing the Dutch wordnet hierarchy top-down, revising all synsets with a large number of equivalence relations or low-scoring equivalence relations to English Wordnet, and clarifying the synset-type relations for large sets of co-hyponyms. This will result in more consistency and an measurable overall improvement of the database. Details about these strategies can be found in *D03 Top level ontology, relation constraints and assignments*.

4.2 Integration of multiword units

We will make a start with the inclusion and description of fixed multiword units. In RBN all combinations of words – ranging from free, illustrative examples to fixed multiword units like idioms and proverbs – are given within the microstructure of a particular word contained in the expression. Those combinations which have a reduced semantic transparency and a reduced lack of compositionally are regarded as a semantic unit, are introduced in the macrostructure, and linked to a synset and via the synset to the ontology.
The following steps are carried out:

1. selection of fixed multi word units

In RBN the examples have been distinguished in types like lexical collocations, grammatical collocations, proverbs, pragmatic formulas, etc.[1] On the basis of these example types, we made a selection of candidate multiword expressions consisting of 9500 idioms, 350 proverbs and 5000 lexical collocations. For each candidate mwe, a new multiword lexical unit (MWLU), will be created.

2. The xsd of the Cornetto database is extended to store the MWLU's
3. 50 candidate MWLU's will be manually aligned to a synset and to SUMO.

[1] More information in 'Handleiding RBN2.0' (2005), Martin e.a. , to be found at www.tst.inl.nl/RBN